DOCUMENT RESUME

ED 075 517 UD 013 412

AUTHOR	O'Connor, Edward; Klein, Stephen
TITLE	A Critique of the Report by Irv Garfinkel and Edward M. Gramlich entitled "A Statistical Analysis of the OEO Experiment in Performance Contracting."
PUB DATE	Feb 73
NOTE	32p.; paper presented at the American Educational Research Association annual meeting, New Orleans, La., February 1973
AVAILABLE FROM	Dr. Edward O'Connor, Div. of Analytical Studies, Educational Testing Service, Princeton, N.J. 08540 (no charge)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Compensatory Education Programs; Educational Planning; Educational Policy; *Evaluation Methods; *Evaluation Needs; Experiments; Matched Groups; Multiple Regression Analysis; *Performance Contracts; Program Development; *Program Evaluation; Research Methodology; Statistical Analysis

ABSTRACT

This paper shows that Garfinkel and Gramlich's conclusion about the average performance of the contractors is not supported by the data. The discussion focuses on three issues: (1) The study was conducted in such a way that performance contracting schools and the comparison schools were generally not comparable in terms of initial achievement, socioeconomic status, and a host of other demographic and social variables that may have influenced the results. Since the contractors were generally assigned to the schools with the lower achieving students, who also had lower S.E.S. and family income, the effect of the noncomparability of the comparison schools was to negatively bias the estimated effect of performance contracting. (2) The statistical adjustments used by Garfinkel and Gramlich were inadequate to offset these biases. (3) The data base was open to a wide variety of potential biases that were not assessed; e.g., the testing conditions (at some schools) were terrible, the tests may have been highly speeded and thus not measuring reading and mathematics ability per se, and there was no control on the control group in that they may have been trying harder to outdo the experimentals. (Author/JM)

A Critique of the Report
by Irv Garfinkel and Edward M. Gramlich
entitled

A STATISTICAL ANALYSIS OF THE OEO EXPERIMENT IN PERFORMANCE CONTRACTING

This critique was prepared by Dr. Edward O'Connor of Educational Testing Service and by Dr. Stephen Klein of UCLA and Educational Evaluation Associates. Copies may be obtained by writing:

Dr. Edward O'Connor
Division of Analytical Studies
Educational Testing Service
Princeton, New Jersey 08540

[illegible]

A Critique of the Report
by Irv Garfinkel and Edward M. Gramlich
entitled
A STATISTICAL ANALYSIS OF THE OEO EXPERIMENT IN PERFORMANCE CONTRACTING

Edward F. O'Connor, Jr.
Educational Testing Service,

and

Stephen P. Klein
University of California, Los Angeles

The OEO experiment "was designed to test whether private companies operating with their existing technologies on an incentive basis could provide better remedial education to poor students than the normal schools." (G & G, p.1) It was not expected that all private companies could outperform all public schools under all circumstances. On the contrary, different companies were expected to use different methods under different circumstances and to achieve varying degrees of success.

The central issue in the OEO experiment was whether performance contracting was worth pursuing, i.e., does it appear likely that some contractors can significantly out-perform some public schools? There is considerable evidence in the Garfinkel and Gramlich report that some contractors did indeed out-perform the public schools at certain sites. Despite this evidence, Garfinkel and Gramlich concluded that "performance contractors who participated in the experiment do not currently have the capability of bringing about any great improvement in the educational status of disadvantaged children (p. 27)." The basis for their conclusion was that their analysis indicated that the <u>average</u> performance of the contractors was very similar to the average performance of the public schools. In other words, they are implying a conclusion about the individual contractors on the basis of the group's performance as a whole--averaging the successes and the failures. Whether contractors as a whole out-performed public schools was essentially a non-question: it was anticipated that some contractors would fail. The essence of the performance contract concept is that the performance contractors who consistently fail will have to change their methods or go out of business (as many of them have). The major issue was whether there is evidence that some performance contractors can out-perform some public schools. The OEO experiment was not adequately designed to ask this question, but there is considerable evidence in the data that some contractors consistently out-performed the public schools. We will show in this paper that Garfinkel and Gramlich's conclusion about the average performance of the contractors is not supported by the data. This discussion will focus on three issues.

1. The study was conducted in such a way that performance contracting schools and the comparison schools were generally not comparable in terms of initial achievement, socioeconomic status, and a host of other demographic and social variables that may have influenced the results. Since the contractors were generally assigned to the schools with the lower achieving students, who also had lower SES and family income, the effect of the non-comparability (or mis-matching) of the comparison schools was to negatively bias the estimated effect of performance contracting.
2. The statistical adjustments used by Garfinkel and Gramlich were inadequate to offset these biases.
3. The data base was open to a wide variety of potential biases that were not assessed; e.g., the testing conditions (at some

schools) were terrible, the tests may have been highly speeded and thus not measuring reading and mathematics ability per se, and there was no control on the control group in that they may have been trying harder to outdo the experimentals (called the "John Henry Effect," Saretsky, 1972.)

Inadequacies in the Experimental Design

In an ideal experiment, the random assignment of students to the experimental and control groups allows the researcher to state with a high degree of confidence that the posttest differences between the experimental and control groups are due to something which occurred between the time of the random assignment and the posttest. But in many educational settings, the random assignment of students is not possible and alternative strategies must be employed. One alternative approach is to randomly assign comparable schools to the experimental and control groups. Campbell and Stanley (1963) refer to this approach as a "quasi-experimental" design. Another alternative is to choose the experimental schools and then find comparable "control" schools. Campbell and Stanley refer to this approach as a "pre-experimental" design because it has questionable validity under even the best of circumstances. This latter approach was the one used in the OEO study despite very apparent and unfavorable circumstances, namely, the fact that the control schools had more able and affluent students.

Why was the pre-experimental design chosen? The details in this particular case are probably quite complex and involve a variety of factors, personalities, and policy decisions. It is interesting to note, however, that most federal evaluation efforts have put the less able and affluent in the experimental group, e.g., Headstart and Title I. Perhaps the decision-makers feel that the treatment will be beneficial and it should go to the students who appear to need it and can profit from it most. Unfortunately, this very method of assignment ensures that the comparison schools and students are not comparable in a variety of known and unknown ways. Thus, the data are often "adjusted" statistically in an attempt to control for the non-comparability. For a variety of technical reasons that will be discussed later in this report, the initial differences between the treatment groups are under-adjusted. Since the experimental treatment has generally been assigned to the lower achieving schools, the consistent result has been a potential under-estimate of the effect of social action educational programs, such as Headstart, Title I, and the OEO Performance Contracting Study. It must be reiterated, however, that these attempts to make statistical adjustments were initiated only because the researchers failed to use the more powerful research designs available to them. In the OEO study, for example, a much more powerful design would have been to randomly assign schools to one of the two treatments at each site. This would have resulted in essentially the same initial position of the two groups (experimental and control) for at least the aggregate sample. Such a design would have essentially eliminated many of the statistical and logical problems which will be discussed in connection with OEO's analyses.

In addition to the non-random assignment of school, there were other inadequacies in carrying out the experimental design. Different criteria were used to select the experimental and control school. "Generally speaking, the most deficient school or schools were selected as the experimental school(s) and the next most deficient as the control schools (Ray, 1972, p. 8)."

Within the experimental and control schools, different criteria were sometimes used to select students for the experimental and control groups. It is likely that in some cases less care was employed to select the control students since they were not going to benefit from the experimental program. Attrition was handled differently in the experimental and control groups. "Thus, of the 106 site/grade/subject area combinations at the secondary level, the control group has a higher grade equivalency entry in 84, or approximately 80 percent of these combinations (Ray, 1972, p. 34)."

Despite the fact that median income and other family background data was needed as covariates for some of the statistical adjustment procedures, the overall survey return for the parents' questionnaire was less than fifty percent, varying from zero percent at some schools to an estimate of over one hundred percent at a Dallas eight grade class. "This can be only accounted for in terms of Dallas eight grade adding more control students after the initial master list was created. The situation at Dallas was probably replicated at other sites; hence response rates should be interpreted cautiously" (Ray, 1972, Appendix B). It must be added that not only the return rate but data themselves must be interpreted cautiously.

Racial data was not collected at four of the sites. For ten of the fourteen sites where racial data was collected, the percent of whites in the control group exceeded the percentage of whites in the experimental groups (in one school 37% higher).

Median income was also used in the adjustment methods. The response rate varied from lows of zero percent for the Rockland control group and eight percent at the Philadelphia control group to a high of 98 percent for the Taft control group. Median income varied by better than two to one at some sites (Anchorage: second grade experimental group, $8,000; second grade control group $16,950; McComb: third grade experimental group, $4,375; third grade control group, $9,500).

Inadequacies in the Statistical Analysis

In most evaluation studies in which matching or statistical adjustments have been used to control for pre-existing differences in the treatment groups, there has been no outside criterion which could be used to validate the matching or adjustment procedures. For example, one can choose to believe (Circirelli, 1970; Evans & Schiller, 1970) or not believe (Campbell & Erlebacher, 1970) that the matching procedures were adequate in the Westinghouse-Ohio evaluation of Headstart. The data needed to validate the procedures simply was not available. Fortunately in the OEO study, there is an outside criterion which can be used to examine the validity of the adjustment procedures. This opportunity came about as a result of the fact that classes with each school differed greatly in their pretest scores. Consequently some experimental classes were superior to their control counterparts even though the experimental school as a whole scored below its control school. This meant that for every kind of statistical adjustment employed, a test could be run to examine its potential efficacy. Unfortunately, this is a one-sided test in that it can only say whether a particular adjustment is biased in a given manner; it cannot say that the adjustment is unbiased. Furthermore the test is only sensitive to relatively large biases. The nature of this test is illustrated by the following table:

"Adjusted" Posttest Results

		Most favorable for experimental group	Middle third	Least favorable for experimental group
Initial status difference between groups	Most favorable for experimental group			
	Middle third			
	Least favorable for experimental group			

This table compares the initial status of the sites with their final status on the "adjusted" posttest treatment effects. For each grade the eighteen sites were ranked according to the difference between the experimental and group on one initial status variable, either the reading or mathematics pretest, or median family income, or percent white, with the largest difference in favor of the experimental group listed first and the largest difference favoring the control group listed last.* The sites were then split into thirds, the third most favorable to the experimental group, the middle third, and the third most favorable to the control group. The same procedure was followed for each of Garfinkel and Gramlich's five treatment effect estimates and the Battelle estimate. A three by three contingency table was prepared with the initial status differences determining the rows and the "adjusted" posttest differences determining the columns. The tables were prepared by

*There was a small difference in the samples used by Battelle and by OEO for their analyses. These differences were small enough to be ignored in this analysis.

summing across all six grades in order to obtain a large enough sample size to test for significance.** A significant chi-square means that the adjusted posttest differences are not independent of initial status and hence are biased. A summary of the chi-squares are presented in the Table 2 in the appendix. The six estimates of the treatment effect for reading are designated R_1 through R_6 and the six estimates for math are designated as M_1 through M_6.

The first set of chi-squares compares each twelve treatment effect estimates with initial status on the pretest of the same subject. All of these chi-squares were non-significant. The second set compares each of the twelve estimates with the opposite pretest. Four of these chi-squares are significant at the .05 level or better. The third set compares treatment effect estimates with median family income. Five of the chi-squares were significant, all of them in mathematics. The fourth set compares the treatment effect estimates with percent white. Two of these chi-squares were significant, both of them in reading.

The tables with significant chi squares are presented in Tables 4a through 4k in the appendix. All of these tables show a tendency for the treatment effect estimates to be positively related to initial status, that is, the groups with the largest estimated treatment effect favoring the experimental group tended to be those groups with the largest initial status difference favoring the experimental group. This positive correlation between

**The first graders took only a single pretest combining reading and math readiness. Consequently their data was excluded from the "appropriate pretest" chi square tables. Otherwise all six grades are included in each table.

initial status and estimated treatment effect means the estimates are biased. Five of the six methods for estimating the treatment effects showed at least one significant chi-square. The one exception, surprisingly enough, was the unadjusted raw score mean gain difference, an estimate which we and Garfinkel and Gramlich consider to be biased on a priori grounds, which will be discussed later.

The six methods to estimate the treatment effect are described in Table I.

Why the Statistical Adjustments Failed

The simplest explanation for the failure of the statistical adjustments may be the most profound: the schools were simply different. They were different on the achievement pretests, on parental income, on parental education, and on a variety of other known and unknown variables. There is simply no known statistical procedure that can be counted on to make the appropriate adjustment in such cases (Lord, 1967, 1969). In this particular case, we have demonstrated that the adjustments were not appropriate. All of the statistical adjustments rest on a series of assumptions and in this section we will demonstrate that these assumptions were not met in the OEO study. The effect of failing to meet these assumptions was to systematically underestimate the effect of performance contracting.

All the adjustment procedures are essentially attempts to predict what the difference between the treatment groups would have been in the absence of the treatment. If the predicted difference and the obtained difference are about the same, one concludes that there was no treatment effect. On the other hand, if the predicted and the obtained differences are sufficiently discrepant, one concludes that the discrepancy is due to the effect of the treatment. In short, the adjustment procedures require certain assumptions about the predictability of the academic growth of the treatment groups (i.e., what would have happened if what did happen had not happened).

The simplest assumption is that the experimental and control groups would have grown by the same number of raw score points if they had received the same treatment. Garfinkel and Gramlich present the mean gain difference on line 1 of Table III. For example, on the reading test the first grade experimental group gained one point more than the first grade control group did. If we could assume that the two groups would have gained the same amount in the absence of the experimental treatment (performance contracting), the one point difference could be attributed to performance contracting. But the assumption of equal gains seems unlikely in light of the data indicating that the experimental group had lower pretest scores, family income, and SES than did students in the control group. There is sufficient evidence in the OEO report and press release, as well as in other research (Coleman, 1966; Hubert, 1972) that students who start out low on these dimensions have a slower rate of growth in skill development than do students who are more able and/or come from more affluent homes. It would be expected, therefore, that the experimental group would gain less in test scores than the control group if there were no effect of performance contracting. Thus, the mean unadjusted gain differences are negatively biased estimates of the effect of performance contracting, i.e., they underestimate its impact. It is interesting to note, therefore, that since the unadjusted mean gains were essentially the same for the two groups, there is actually some evidence to support the contention that performance contracting had a generally positive impact on scores.

Garfinkel and Gramlich also concluded that the mean gain differences were negatively biased estimates of the experimental effect. On line 4 of Table III they presented the "adjusted" mean gain differences. This adjustment process assumes that the pretest-posttest relationship is the same between groups and within groups. This assumption holds reasonably well for randomized groups, but is not necessarily true for non-randomized groups. In fact, the between-group and within-group relationships may even have different signs (Robinson, 1950). The assumption that the between-group and within-group relationships would have been identical in the absence of the treatment effect is known as the "ecological fallacy." This problem has usually been discussed in terms of the difficulty of inferring individual behavior from the behavior of group averages, but the general principle is the same: variables do not necessarily affect total group means in the same way that they affect subgroups or individuals within this total group. For a fuller treatment of this problem, see Selvin (1958), Cartwright (1969), and Hannan (1971).

Garfinkel and Gramlich presented three regression estimates of the treatment effect on lines 2, 3, and 5 of Table III. These estimates depend on the assumption that the between-groups and within-group regression slopes would have been identical in the absence of any treatment effect. This is another example of the ecological fallacy. O'Connor (1972) demonstrates that estimates of school treatment effects can be biased when they are based on the within-group regression slopes.

Even if we were to ignore the ecological fallacy (and assume that the within-group and between-groups regression slopes would have been identical), the Garfinkel and Gramlich analysis would still be invalid for the following three reasons:

1. Specification error
2. Errors of measurement
3. Heterogeniety of the within-group regression slopes.

The first and third reasons apply to all five of the estimates by Garfinkel and Gramlich and to the Battelle estimate and the second reason applies to the second and third estimate presented by Garfinkel and Gramlich and to the Battele estimate.

"Specification error" occurs when the treatment groups are different on one or more variables which are correlated with achievement and which are not used in adjustment equations. For example, we know that parental income is related to achievement and that the experimental and control groups were different on average parental income. If parental income, initial achievement (the pretest scores), and the treatment (performance contracting) were the only variables that affected final achievement, the correct mathematical model would be:

Final Achievement = initial achievement effect + parental income effect + the treatment effect.

In such a case, the following model would be an oversimplification of reality and would produce a biased estimate of the treatment effect:

Final Achievement = initial achievement effect + the estimated treatment effect.

The estimated treatment effect in the second equation would equal the actual treatment effect plus part of the parental income effect. In other words, the second equation gives the treatment credit for part of the parental income effect. The bias resulting from the failure to include (specify) parental-income in the equation is one form of "specification error."

Garfinkel and Gramlich's first, second, fourth and fifth estimates are based on equations which include only the initial achievement scores, not SES, parental income, parental education, or other variables on which the experimental and control groups differ. Since the control group was initially higher on these variables, the probable result of the specification error was to underestimate the effect of performance contracting.

The estimates on line 3 were based on an equation which included additional variables such as family income, parental education, race, sex, and age. However, the equation did not correct for errors of measurement and did not include all of the variables which might be related to achievement (e.g., Coleman Report, 1966). Further, the data were too incomplete (55% response rate for family income, p. 4) to place a high reliance on the results (parents who return questionnaires are likely to be different than those who do not).

The estimates on lines 2 and 3 were based on equations which did not correct for errors of measurement in the pretest data (initial achievement scores, parental income, etc.). (Note: No correction for errors of measurement was required for the estimates on line 1. However, these estimates are likely to be biased by specification error.) Since the control group had higher initial scores, the probable result of failing to make this correction was to underestimate the effect of performance contracting.

The estimate on lines 4 and 5 are corrected for errors of measurement in the pretest achievement scores but they are not adjusted for the effect of the variables such as parental income and education which are known to influence achievement. Although the estimates are labeled differently, both essentially regression coefficients adjusted for errors of measurement according to a similar set of assumptions. The probable result of the specification errors in these estimates is to again underestimate the effect of performance contracting.

Another way to look at estimates which are "adjusted" for the effect of errors of measurement is to realize that the correction, at best, produces the same estimate that one would have obtained with a perfectly reliable measure of initial status. Lord (1967) demonstrated that even with a perfectly reliable measure of initial status, the estimated treatment effects are not a reliable indicator of the actual treatment effect when the groups are different on the pretest measures.

All of the estimates presented by Garfinkel and Gramlich are based on the further assumption that the experimental and control groups have the same within-group regression slopes. Table V (G & G) shows that experimental and control groups have significantly different regression slopes for at least first grade reading and mathematics, and for 7th and 8th grade mathematics (i.e., 4 of the 12 grade subject combinations). In addition, the true score regression slopes presented in Table II appear to be substantially different for 2nd grade reading and 3rd grade reading and mathematics, although no test of statistical significance was performed. Without a re-analysis of the data, it is not possible to state whether these differences in the regression slopes biased the Garfinkel and Gramlich estimates of the treatment effect. Because of the other inadequacies of the experimental design, it is probably not worthwhile to pursue this re-analysis.

The differences in the regressions slopes have practical implications as well, despite Garfinkel and Gramlich's statement "that the differences in slope between experimental and control students is very slight, being statistically significant in only a few cases and never amounting to much quantitatively" (p. 20). In the four cases where the slopes are significantly different, the experimental group has the smaller regression slope. This indicates that the performance contractors were relatively more successful with low initial achievement students than were the public schools. There are two alternative explanations for this finding: (1) the performance contractors were successfully concentrating their efforts on the low initial achievement students, their designated target, or (2) the difference in the slopes is simply another indication of how really different the experimental and control groups were initially in terms of their patterns of academic growth. Because the schools were not randomly assigned to experimental and control groups, there is no way to definitely choose between these two alternative explanations.

The Battelle estimates, R_6 and M_6, are perhaps the most interesting because they attempted to take in consideration possible differences in the within-group regression slopes. Figures 1, 2, and 3 illustrate this approach. The Battelle approach compares the two regression slopes at the mean of the combined groups. In the Figure 1a, the two regression slopes are identical and hence the estimated treatment effect is zero at the combined group mean and at every other point.

Figure 2 presents the same means but this time the experimental group has a steeper regression slope than the control group. Under these circumstances, the Battelle estimates of the treatment effects is positive. Figure 3 presents the same means but in this example, the experimental group has a flatter regression slope and the Battelle estimate of the treatment effect is negative. Note that in all three cases, the estimated treatment effect at the experimental group pretest mean is zero. In other words, the experimental treatment effect was not estimated at the pretest mean of the group that the performance contractors actually had to work with, but at the pretest mean of some hypothetical combined group.

This appears to be an unreasonable way to evaluate any program. Furthermore, if performance contracting is to be used to upgrade the performance of the lowest achieving students, it should concentrate on the lowest achieving students with the result that the regression slope is relatively flat. In contrast, the Battelle approach assigns a positive treatment effect to the performance contractors with a steep slope and a negative treatment effect to the performance contractor with the flatter slope, exactly the opposite of what is socially desirable.

Intercorrelations of the Various Estimates

We can hypothesize a true model which together with error-free data would give us the true treatment effect for each site. All of the estimates derived from the six methods discussed earlier will deviate to some extent from the true estimates because of specification errors, sampling errors, and measurement errors. By examining the intercorrelations of the various estimates, we can get a very rough indication of how well these estimates might be correlated with the true effects. This is somewhat analogous to parallel-forms reliability coefficients.

Tables 3a and 3b present these intercorrelations separately for each grade and the mean over all six grades. The correlations range from .99 to -.01 for reading and from .995 to .21 for mathematics. Any correlation below .75 can be considered significantly below .90, a minimum acceptable level of reliability for the standardized achievement tests (p = .05, one-tailed test). Slightly over half of the correlations and half of the means are below .75. In short, the estimates are poorly correlated and cannot form the basis for a valid determination of the true effects unless a clear case that can be made one of the estimates is more highly correlated with the true effect than the others are. To our knowledge, no such case can be made with these data. It is worth noting that although we have been discussing the estimates of treatment effects at the individual sites, the same arguments hold for the estimates of the treatment effects over all sites. Since the errors we have discussed are systematic errors, they affect the overall estimates as well as the site estimates. Consequently, no greater confidence can be placed in the overall estimates than in site estimates.

For the purpose of comparison we have presented in Table 3c the correlations of the Reading estimates with the Mathematics estimates holding the method constant. These range from .89 to -.04. Unfortunately we can not determine from this data whether the positive correlations are due primarily to the contractors having equal success with both subjects or to shared biases in the procedures.

Summary and Conclusions

Even if the schools had been randomly assigned to the experimental and control groups and appropriate statistical methods had been devised, it is likely that the inadequacies in the assignment of students and the data collection would have rendered the data uninterpretable. Most of these inadequacies can be attributed to OEO's unwillingness to allow sufficient time for the proper planning of this experiment.

There has been a consistent theme across many federal evaluation efforts for the experimental groups to contain students who are lower scoring and less afluent than the control students. The experimental program are assigned to the schools and the students who appear to have the greatest need for the program. The evaluators then look for "comparable" control schools. Unfortunately, this method of assignment ensues that the comparison schools and students are not comparable in a variety of known and unknown ways. Consequently, the data must be "adjusted" statistically in an attempt to control for the non-comparability. It is clear from the OEO data and from a number of theoretical articles that these adjustments cannot compensate for deficiencies in the experimental design. Perhaps it can be argued that in the case of large scale national programs such as Title I, there is no politically feasible alternative way to conduct the evaluation. Whatever the merits of that argument, it seems clear that in small pilot programs such as the OEO experiment, it is feasible to randomly assign schools to the experimental and control conditions.

We are also troubled by the persistence of the federal evaluations to make arbitrary summative evaluations about highly diverse programs. Terms like performance contracting, Head Start, Title I, and compensatory reading do not define experimental treatments. Undoubtedly, there are some performance contractors who could out-perform some schools under some circumstances. The question was which contractors could out-perform which schools under what conditions? The OEO experiment in performance contracting was neither designed nor analysed to adequately answer that question.

Table 1

A description of the six methods used by OEO and by Battelle to Estimate the Treatment Effects.

R_1 and M_1 : simply the difference between the mean gain of the experimental group and the mean gain of the control group.

R_2 and M_2 : a regression model using same subject pretest as the only covariate and allowing for non-linearities.

R_3 and M_3 : a regression model using same subject pretest and "a vector of other independent variables including average family income, education of parents, race sex, and age" as covariates and allowing for non-linearities in the relationship with pretest.

R_4 and M_4 : R_1 and M_1 adjusted for "bias."

R_5 and M_5 : R2 and M_2 adjusted for "bias."

R_6 and M_6 : the Battelle estimate derived from a regression model which used the same subject pretest as the only covariate, allowing for differences in the within-group regression slopes. This estimate is discussed in more detail in the text.

	R_1	R_2	R_3	R_4	R_5	R_6	M_1	M_2	M_3	M_4	M_5	M_6
Same Pretest	4.90	8.06	7.81	8.62	3.70	9.23	2.74	3.25	1.15	4.29	9.34	3.51
Opposite Pretest	3.66	9.75 p=.05	12.08 p=.05	8.52	2.75	8.41	2.83	8.27	14.51 p=.01	0.59	2.23	15.61 p=.01
Income	5.05	3.69	4.02	0.35	1.24	4.75	5.14	9.92 p=.05	19.21 p=.001	9.90 p=.05	12.21 p=.05	14.24 p=.01
Percent White	2.87	10.17 p=.05	10.53 p=.05	7.11	1.78	6.62	2.28	0.05	2.57	9.33	1.78	4.96

Table 2 Chi-square comparing estimated treatment effects with initial status.

	Gr	R1	R2	R3	R4	R5	R6
R2	1	.39 (15)					
	2	.88 (18)					
	3	.87 (18)					
	7	.83 (17)					
	8	.96 (18)					
	9	.87 (17)					
	M	.80					
R3	1	.40 (17)	.99 (15)				
	2	.36 (17)	.55 (17)				
	3	.85 (18)	.97 (18)				
	7	.84 (17)	.99 (18)				
	8	.92 (18)	.96 (18)				
	9	.78 (17)	.71 (17)				
	M	.69	.86				
R4	1	.68 (17)	.65 (15)	.64 (17)			
	2	.95 (18)	.73 (18)	.29 (17)			
	3	.74 (18)	.69 (18)	.71 (18)			
	7	.89 (16)	.55 (17)	.57 (17)			
	8	.98 (18)	.91 (18)	.84 (18)			
	9	.88 (17)	.58 (17)	.82 (17)			
	M	.85	.68	.64			
R5	1	.41 (17)	.36 (15)	.34 (17)	.44 (17)		
	2	.66 (18)	.62 (18)	.38 (17)	.73 (18)		
	3	.03 (18)	.04 (18)	.10 (18)	.34 (18)		
	7	.36 (16)	.46 (17)	.45 (17)	.09 (17)		
	8	.61 (18)	.47 (18)	.50 (18)	.62 (18)		
	9	.55 (17)	.69 (17)	.34 (17)	.24 (17)		
	M	.44	.44	.35	.41		
R6	1	.40 (17)	.97 (15)	.95 (17)	.71 (17)	.43 (17)	
	2	.90 (18)	.95 (18)	.43 (17)	.75 (18)	.53 (18)	
	3	.92 (18)	.90 (18)	.88 (18)	.70 (18)	-.01 (18)	
	7	.81 (17)	.90 (18)	.90 (18)	.57 (17)	.36 (17)	
	8	.96 (18)	.98 (18)	.95 (18)	.91 (18)	.49 (18)	
	9	.82 (17)	.68 (17)	.90 (17)	.90 (17)	.17 (17)	
	M	.80	.90	.84	.76	.33	

Table 3a: Intercorrelations of the Six Estimates of the Treatment Effcts for Reading. The means are listed below each column and the number of paired sites are present in the parentheses.

	Gr	M1	M2	M3	M4	M5	M6
M2	1	.69 (15)					
	2	.34 (18)					
	3	.96 (18)					
	7	.96 (17)					
	8	.92 (18)					
	9	.80 (17)					
	M	.78					
M3	1	.70 (15)	.995 (16)				
	2	.58 (18)	.21 (18)				
	3	.96 (18)	.95 (18)				
	7	.91 (18)	.94 (17)				
	8	.89 (18)	.97 (18)				
	9	.73 (17)	.48 (17)				
	M	.80	.69				
M4	1	.79 (16)	.93 (16)	.93 (17)			
	2	.93 (18)	.29 (18)	.52 (18)			
	3	.71 (18)	.65 (18)	.69 (18)			
	7	.92 (17)	.79 (16)	.79 (17)			
	8	.88 (18)	.80 (18)	.79 (18)			
	9	.76 (17)	.26 (17)	.82 (17)			
	M	.83	.62	.76			
M5	1	.45 (16)	.66 (16)	.65 (17)	.75 (17)		
	2	.59 (18)	.60 (18)	.17 (18)	.49 (18)		
	3	.71 (18)	.71 (18)	.70 (18)	.51 (18)		
	7	.53 (17)	.58 (16)	.51 (17)	.60 (17)		
	8	.83 (18)	.62 (18)	.58 (18)	.81 (18)		
	9	.50 (17)	.17 (17)	.56 (17)	.61 (17)		
	M	.60	.56	.53	.63		
M6	1	.64 (16)	.97 (16)	.96 (17)	.92 (17)	.68 (17)	
	2	.83 (18)	.31 (18)	.60 (18)	.78 (18)	.32 (18)	
	3	.99 (18)	.97 (18)	.98 (18)	.70 (18)	.75 (18)	
	7	.96 (18)	.97 (17)	.96 (18)	.80 (17)	.46 (17)	
	8	.97 (18)	.94 (18)	.92 (18)	.81 (18)	.74 (18)	
	9	.82 (17)	.41 (17)	.92 (17)	.96 (17)	.63 (17)	
	M	.87	.76	.89	.83	.60	

Table 3b: Intercorrelations of Six Estimates of the Treatment Effects for Mathematics. The means are listed below each column and the number of paired sites are present in the parentheses.

Gr	R1/M1	R2/M2	R3/M3	R4/M4	R5/M5	R6/M6
1	.63 (16)	.72 (14)	.71 (17)	.57 (17)	.89 (17)	.66 (17)
2	.63 (18)	.49 (18)	.39 (17)	.61 (18)	.45 (18)	.69 (18)
3	.69 (18)	.84 (18)	.86 (18)	.35 (18)	.11 (18)	.74 (18)
7	.14 (17)	.49 (17)	.50 (18)	.03 (17)	.37 (17)	.38 (18)
8	.43 (18)	.36 (18)	.46 (18)	.30 (18)	.54 (18)	.42 (18)
9	.61 (17)	.69 (17)	.65 (17)	.68 (17)	-.04 (17)	.62 (17)
M	.52	.60	.60	.42	.39	.58

Table 3c: Correlations of the Estimates of the Reading Treatment Effects with the Estimates of the Mathematics Treatment Effects using the same method of analysis.

Table 4a:

	most favorable		least favorable
MF	14	10	6
	11	10	8
LF	5	9	16

Math pretest by R_2

Table 4b:

	most favorable		least favorable
MF	16	8	6
	8	12	8
LF	6	8	16

Math pretest by R_3

Table 4c:

	most favorable		least favorable
MF	15	9	6
	8	14	7
LF	7	6	17

Reading pretest by M_3

Table 4d:

	most favorable		least favorable
MF	18	5	7
	8	11	10
LF	4	13	13

Reading pretest by M_6

Table 4e:

	most favorable		least favorable
MF	12	11	3
	9	6	11
LF	5	9	12

Income by M_2

Table 4f:

	most favorable		least favorable
MF	11	13	2
	8	12	8
LF	7	3	16

Income by M_3

Table 4g:

	most favorable		least favorable
MF	11	10	5
	6	13	8
MF	9	4	13

Income by M_4

Table 4h:

	most favorable		least favorable
MF	13	11	2
	7	9	11
LF	6	7	13

Income by M_5

Table 4i:

	most favorable		least favorable
MF	15	7	4
	8	11	9
LF	3	10	13

Income by M6

Table 4j:

	most favorable		least favorable
MF	11	7	10
	8	13	4
LF	9	5	14

Percent White by R_2

Table 4k:

	most favorable		least favorable
MF	12	8	8
	8	13	5
LF	8	5	15

Percent White by R_3

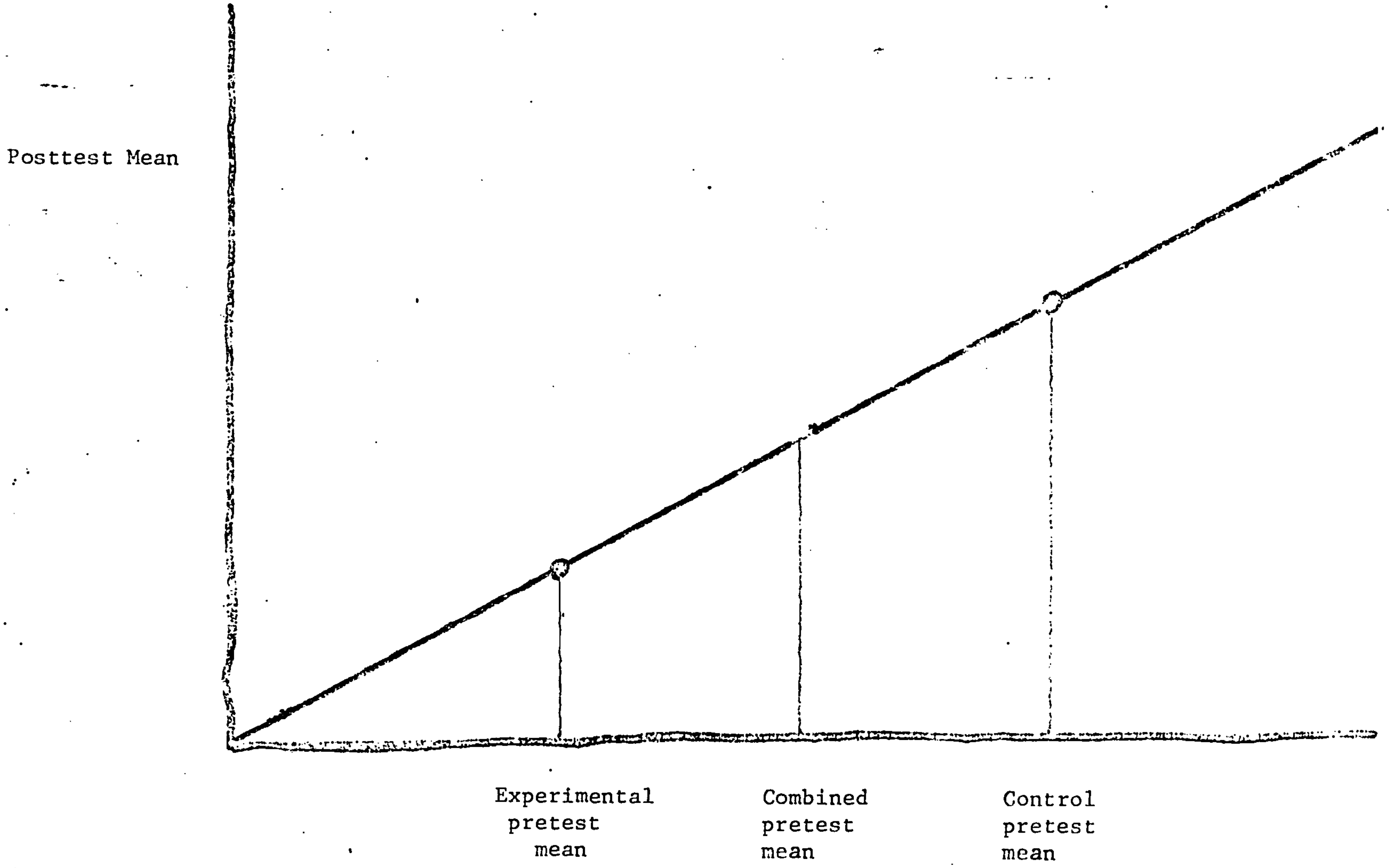

Figure 1 Estimated zero treatment effect with within-group regression slopes the same for both groups

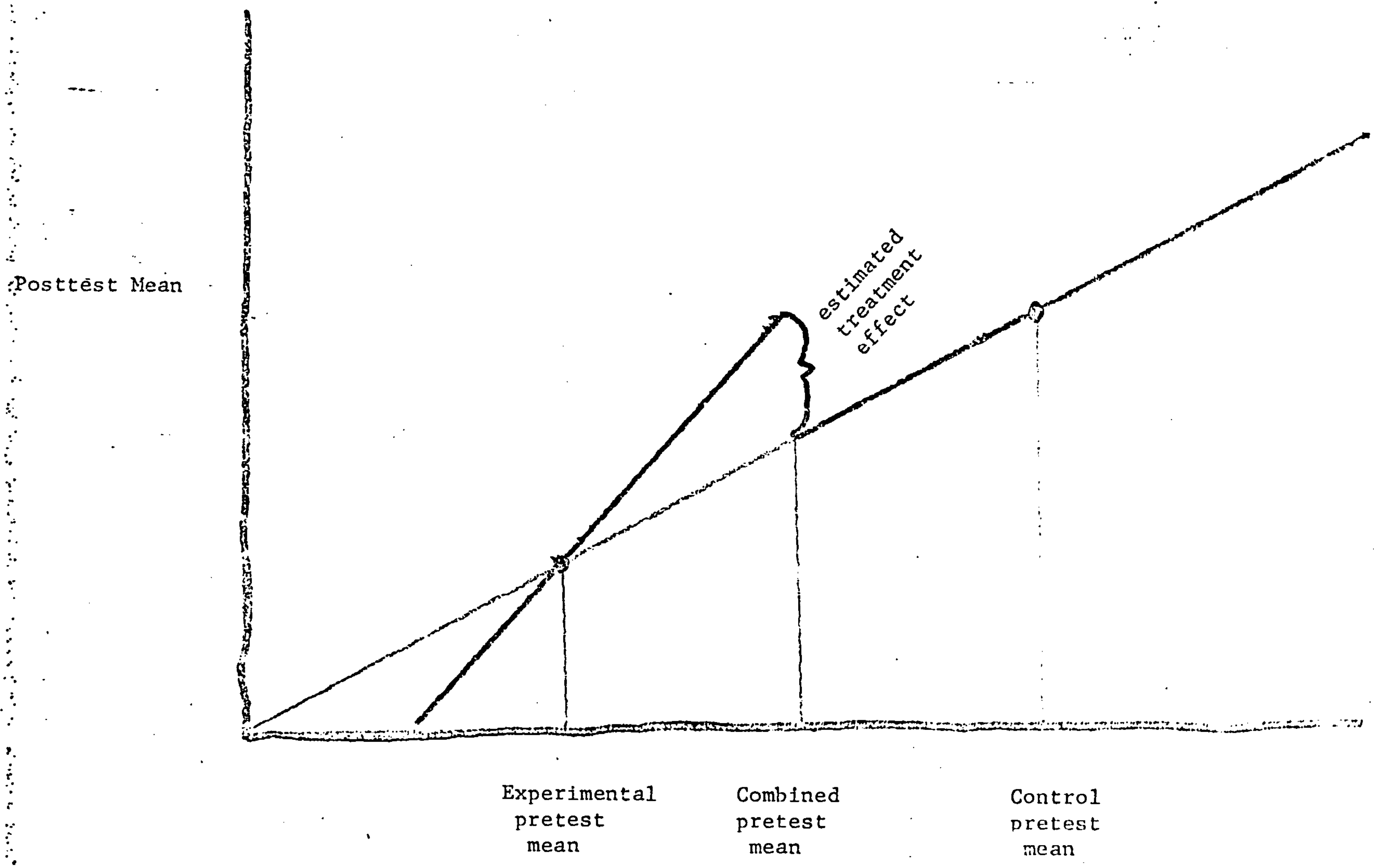

Figure 2 Same means as in 1 but with the experimental group having the steeper regression slope.

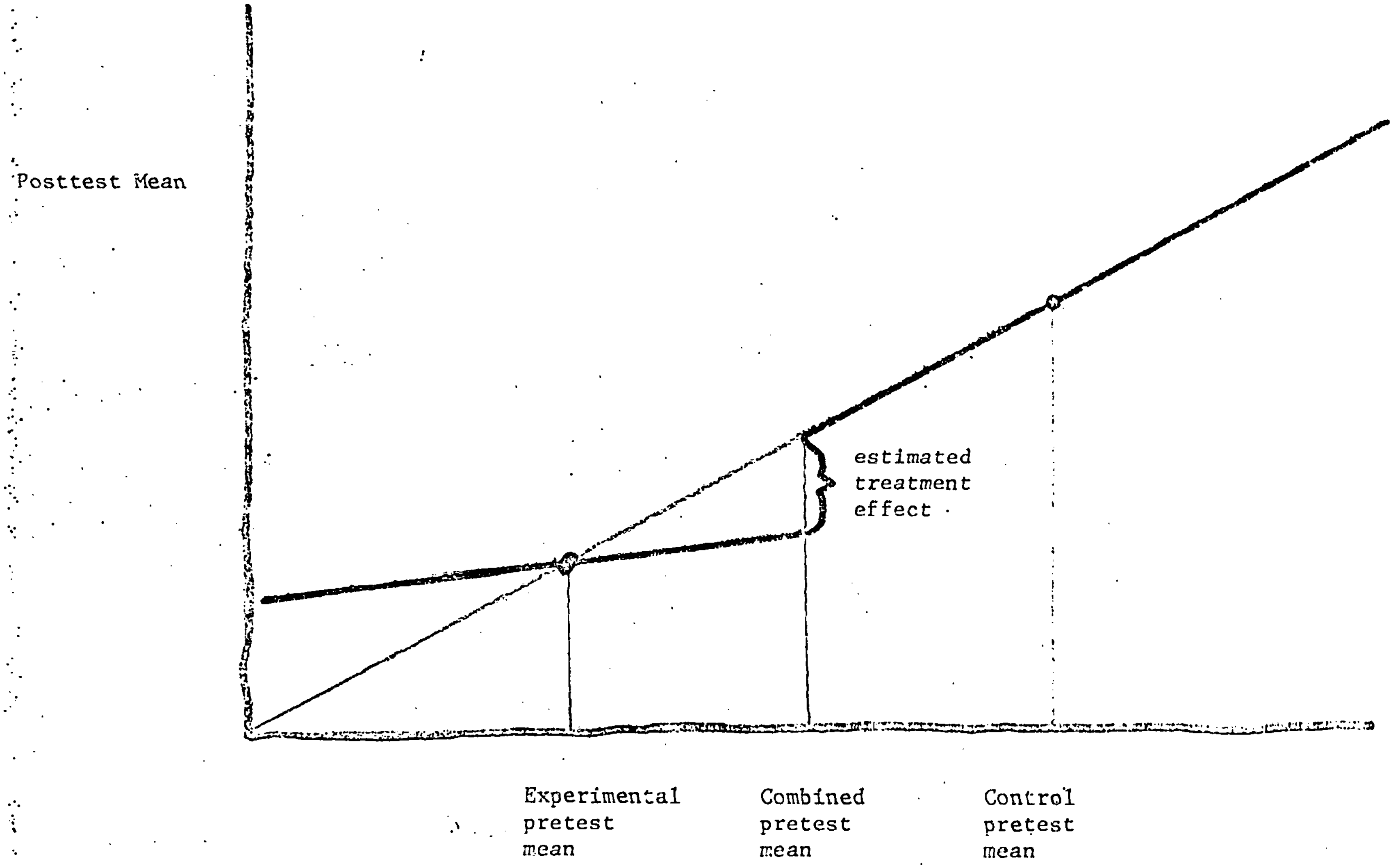

Figure 3 Same means as in 1 and 2 but with the experimental group having a less steep regression slope than the control group.

REFERENCES

Campbell, D. T. & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Helmuth, (ed.), Compensatory education: A national debate, Vol. III of The Disadvantaged Child. New York: Brunner/Mazel, 1970.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Cartwright, D. S. Ecological variables. In E. F. Borgatta (ed.) Sociological Methodology 1969. San Francisco: Josey-Bass, 1969.

Cicarelli, V. G. The relevance of the regression artifact problem to the Westinghouse-Ohio evaluation of Head Start: A reply to Campbell & Erlebacher. In J. Helmuth, (ed.), Compensatory education: A national debate, Vol. III of the Disadvantaged Child. New York: Brunner/Mazel, 1970.

Coleman, J. S. Equality of Educational Opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Evans, J.W. & Schiller, J. How preoccupation with possible regression artifacts can lead to a faulty strategy for the evaluation of social action programs: a reply to Campbell and Erlebacher, in J. Helmuth (ed.).

Garfinkel, I. & Gramlich, E. M. A statistical analysis of the OEO experiment in educational performance contracting. In Office of Economic Opportunity, An experiment in performance contracting, OEO Pamphlet 3400-6, June 1972.

Hannan, M. t. Problems of aggregation. In Blalock, H. M. (ed.). Causal models in the social sciences. Chicago: Aldine-Atherton, 1971.

Hubert, D. The Deluth experience. Saturday Review, May 27, 1972; 55-59.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

Lord, F. M. Statistical adjustments when comparing preexisting groups. Psychological Bulletin, 1969, 72, 336-337.

O'Connor, E. F. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Ray, H. W. Final report on the Office of Economic Opportunity experiment in educational performance contracting. Columbus, Ohio: Battelle Columbus Laboratories, march 14, 1972.

Robinson, W. S. Ecological correlations and the behavior of indivicuals. American Sociological Review, 1950, 15, 351-357.

Saretsky, G. The OEO P.C. Experiment and the John Henry Effect. Phi Delta Kappan, May 1972, 597-581.

Selvin, H. C. Durkheim's Suicide and problems of empirical research. American Journal of Sociology, 1958, 63, 607-620.